

SC|07 IB Network InfiniBand Routing



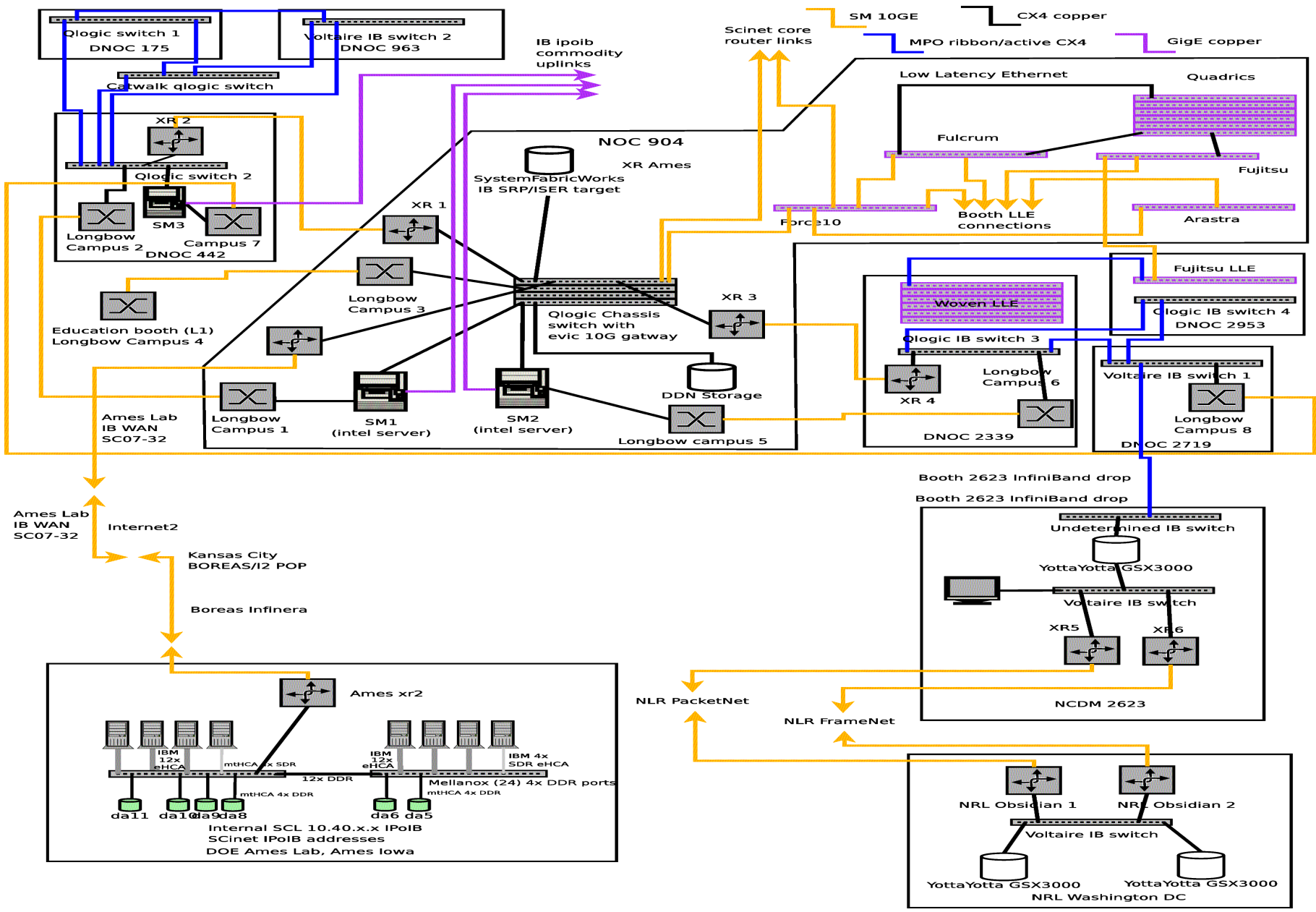
OPEN**FABRICS**
A L L I A N C E

Jason Gunthorpe – Obsidian



Supercomputing '07 SCInet

- Open Fabrics Network including IB and Low Latency Ethernet
- Active optical cabling from Zarlink and Intel for booth connections – both IB and LLETH
- Obsidian Longbows for longer reach IB
- 10GIGE over SMF for LLETH



IB Routing

- First wide scale IB routing demonstration
 - Followup of Obsidian XNET demo at SC|06
- ~5 subnets connected together, two off site in Ames Labs in Iowa and NRL in Washington
- Multivendor – Obsidion Longbow in the core network connected to a SFW software router demonstration.



Success!

- Multiple cross subnet users
- SFW SRP demo
- NRL/YY storage demo
- Global IPoIB over multiple hops



Challenges

- Deploying patched kernels to booths is very hard. Many booths kept with OFED and did not participate
- Finding engaging demos on the network
- Not clear how IPoIB should work with global/local scope

Patch Set used at SC|07


- Work on this started after SC|06
- Many smaller patches for correct GRH/etc are in the mainline
- Various OpenSM patches
- Based off Linus 2.6.22 + opensm GIT
- New patches should be posted to the list next week but may not be inclusion ready

IPoIB Patch (Rolf)

- Currently hard wired to use link local scope for IB multicast groups
 - Prevents crossing routers
- Need this configurable per interface
- Two different approaches now
 - Change the interface scope
 - Create a new child interface with different scope (but same pkey)
- Modifies core kernel code -> complete rebuild

CMA Patches (Sean)

- Non-standard 'permissive LID' signaling relies on the router to select the return path
- Requires modifications to the Active and Passive sides
- Many limitations
- Good enough to test ULPs
- `git://~shefty/rdma-dev.git ib_router` branch



ULP Patches (Chas/Steve)

- SRP and IPoIB CM needed separate patches to work with the CM patches
- CM abstraction seems to be leaking internal details, maybe some cleanup is needed
- RDMA-CM from user space does work



OpenSM (Hal/Sasha/Rolf)

- Existing ROUTER_EXP from Hal
- Multiple prefix route list
- Global IPoIB Multicast group creation
- Various minor GRH fixes (already in git)



IBTA

- LWG is actively working on a router specification.
- Interested IBTA members can participate



Thanks

- Troy Benjegerdes
- Douglas Fuller
- Rupert Dance
- Ben Shapiro
- Hal Rosenstock
- Steve Welch
- Chas Williams
- Rolf Manderscheid
- Sean Hefty